Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal

Ziniu Li
ziniuli@link.cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

January 14, 2021

Mainly based on the COLT 2020 paper:
https://arxiv.org/abs/1906.03804

Outline

Background and Notation

Main Result

Analysis and Proof
- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof
- Proof of Lemma 4
- Proof of Lemma 7

Markov Decison Process

- An infinite-horizon discounted Markov Decision Process (MDP) is a tuple $M = (\mathcal{S}, \mathcal{A}, P_M, r_M, \gamma)$:
  - $\mathcal{S}$ and $\mathcal{A}$ are the finite state and action space, respectively.
  - $P_M(s'|s,a)$ is the transition probability matrix.
  - $r_M : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$ is the deterministic reward function.
  - $\gamma \in (0,1)$ is the discount factor.
- The quality of policy π is measured by value function:

$$\forall s \in \mathcal{S} : \quad V_M^\pi(s) := \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^t r_M\left(s^t, a^t\right) \mid s^0 = s\right]$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad Q_M^\pi(s,a) = r_M(s,a) + \gamma P_M(\cdot \mid s,a)^\top V^\pi$$

Setting of RL: Generative Model

- Generative Oracle: we can directly reset it to any state s_t, after which we can take an action a_t and observe the next state $s_{t+1} \sim P_M(\cdot|s_t, a_t)$ and the reward $r_M(s_t, a_t)$.
  - Compared to the pure MDP problem, we still do not known P_M in advance.
  - Compared to the online RL problem, we can go to any s_t without the planning from an initial state s_0.
  - In particular, we have access to the whole state space and action space (i.e., no exploration issue).
- Example: a perfect simulator (e.g., some video game simulators), where we can load (reset) the state s_t from RAM.
- We focus on the setting of generative model throughout.

Outline

Background and Notation

Main Result

Analysis and Proof

- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof

- Proof of Lemma 4
- Proof of Lemma 7

Algorithm: Model-based Methods

Algorithm 1 Model-based Reinforcement Learning

Input: N.

1: Collect N next states for each state-action pair by calling the generative model.
2: Construct an empirical MDP with $\widehat{P}$:

$$\widehat{P}(s'|s,a) = \frac{\#\text{ times } (s,a) \mapsto s'}{N}.$$

3: $\widehat{\pi} \leftarrow$ Run any planning algorithm on the recovered MDP.

Output: $\widehat{\pi}$.

Main Result

Theorem 1 ([Agarwal et al., 2020]).

Suppose $\delta > 0$ *and* $\epsilon \in (0, \frac{1}{\sqrt{1-\gamma}}]$*. Let* $\widehat{\pi}$ *be any* ϵ_{opt}*-optimal policy for* $\widehat{M}$*, i.e.,* $\left\| \widehat{Q}^{\widehat{\pi}} - \widehat{Q}^* \right\| \leq \epsilon_{\text{opt}}$*. If*

$$N \geq \frac{c\gamma \log \left(c|\mathcal{S}||\mathcal{A}|(1-\gamma)^{-1}\delta^{-1} \right)}{(1-\gamma)^3 \epsilon^2},$$

we have

$$Q^{\widehat{\pi}} \geq Q^* - \epsilon - \frac{5\epsilon_{\text{opt}}}{1-\gamma}$$

with probability at least $1-\delta$*, where* c *is an absolute constant.*

Comparison with Prior Work

Algorithm	Sample Complexity	ϵ-Range	References
Phased Q-Learning	$C\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\epsilon^2}$	$(0,(1-\gamma)^{-1}]$	Kearns and Singh (1999)
Empirical QVI	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^5\epsilon^2}$	$(0,1]$	Azar et al. (2013)
Empirical QVI	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}$	$\left(0,\frac{1}{\sqrt{(1-\gamma)\lvert\mathcal{S}\rvert}}\right]$	Azar et al. (2013)
Randomized Primal-Dual Method	$C\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}$	$(0,(1-\gamma)^{-1}]$	Wang (2017)
Sublinear Randomized Value Iteration	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\cdot \mathrm{poly}\log\epsilon^{-1}$	$(0,1]$	Sidford et al. (2018b)
Variance Reduced QVI	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\cdot \mathrm{poly}\log\epsilon^{-1}$	$(0,1]$	Sidford et al. (2018a)
Empirical MDP + *any* accurate black-box planner	$\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}$	$(0,(1-\gamma)^{-1/2}]$	This work

Table 1: **Sample Complexity to Compute ϵ-Optimal Policies Using the Generative Sampling Model**: Here $|\mathcal{S}|$ is the number of states, $|\mathcal{A}|$ is the number of actions per state, $\gamma \in (0,1)$ is the discount factor, and C is an upper bound on the ergodicity. We ignore $\mathrm{poly}\log(|\mathcal{S}||\mathcal{A}|/\delta/(1-\gamma))$ factors in the sample complexity. Rewards are bounded between 0 and 1.

Outline

Background and Notation

Main Result

Analysis and Proof

- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof

- Proof of Lemma 4
- Proof of Lemma 7

Additional Notation

- Let $P_{s,a}$ denote the vector $P(\cdot|s,a)$ and P^π denote the transition matrix induced by a deterministic policy π.

$$Q^\pi = r + \gamma P V^\pi = r + \gamma P^\pi Q^\pi \quad \text{and} \quad Q^\pi = (I - \gamma P^\pi)^{-1} r.$$

- For the state value function $V \in \mathbb{R}^{\mathcal{S}}$, the variance $\mathrm{Var}_P(V) \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is defined as:

$$\mathrm{Var}_P(V)(s,a) := \mathrm{Var}_{P(\cdot|s,a)}(V), \quad \text{so that} \quad \mathrm{Var}_P(V) = P(V)^2 - (PV)^2,$$

where the squares are applied componentwise.

Additional Notation

- The variance of discounted reward is defined as

$$\Sigma_M^\pi(s,a) := \mathbb{E}\left[\left(\sum_{t=0}^{\infty}\gamma^t r_M\left(s_t,a_t\right) - Q_M^\pi(s,a)\right)^2 \middle| s_0 = s, a_0 = a\right].$$

- The variance satisfies the following Bellman style, self-consistency conditions [Azar et al., 2013, Lemma 6]:

$$\Sigma_M^\pi = \gamma^2 \operatorname{Var}_P\left(V_M^\pi\right) + \gamma^2 P^\pi \Sigma_M^\pi \tag{1}$$

- It is straightforward to verify that $\|\Sigma_M^\pi\|_\infty \leq \gamma^2/(1-\gamma)^2$.

Outline

Background and Notation

Main Result

Analysis and Proof

- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof

- Proof of Lemma 4
- Proof of Lemma 7

Error Decomposition

Lemma 2 (Componentwise bounds).

For any policy π, we have

$$Q^{\pi} - \widehat{Q}^{\pi} = \gamma \left(I - \gamma P^{\pi}\right)^{-1} (P - \widehat{P})\widehat{V}^{\pi}. \tag{2}$$

In addition, we have:

$$Q^{\pi} \geq Q^{\star} - \underbrace{\left\|Q^{\pi} - \widehat{Q}^{\pi}\right\|_{\infty}}_{\text{est.}} - \underbrace{\left\|\widehat{Q}^{\pi} - \widehat{Q}^{\star}\right\|_{\infty}}_{\text{opt.}} - \underbrace{\left\|\widehat{Q}^{\pi^{\star}} - Q^{\star}\right\|_{\infty}}_{\text{est.}}.$$

$\leadsto$ $\left\|\widehat{Q}^{\pi^{\star}} - Q^{\star}\right\|_{\infty}$ is to find an optimal value function and is shown to be minimax optimal [Azar et al., 2013].

Proof of Lemma 2

For any policy π,

$$
\begin{aligned}
Q^{\pi} - \widehat{Q}^{\pi} &= \left(I - \gamma P^{\pi}\right)^{-1} r - \left(I - \gamma \widehat{P}^{\pi}\right)^{-1} r \\
&= \left(I - \gamma P^{\pi}\right)^{-1} \left(\left(I - \gamma \widehat{P}^{\pi}\right) - \left(I - \gamma P^{\pi}\right)\right) \widehat{Q}^{\pi} \\
&= \gamma \left(I - \gamma P^{\pi}\right)^{-1} \left(P^{\pi} - \widehat{P}^{\pi}\right) \widehat{Q}^{\pi} \\
&= \gamma \left(I - \gamma P^{\pi}\right)^{-1} (P - \widehat{P}) \widehat{V}^{\pi}.
\end{aligned}
$$

Proof of Lemma 2

For the second claim,

$$\begin{aligned}
Q^{\pi} - Q^{\star} &= Q^{\pi} - \widehat{Q}^{\star} + \widehat{Q}^{\star} - Q^{\star} \\
&\geq Q^{\pi} - \widehat{Q}^{\star} + \widehat{Q}^{\pi^{\star}} - Q^{\star} \\
&\geq - \left\| Q^{\pi} - \widehat{Q}^{\star} \right\|_{\infty} - \left\| \widehat{Q}^{\pi^{\star}} - Q^{\star} \right\|_{\infty} \\
&\geq - \left\| Q^{\pi} - \widehat{Q}^{\pi} \right\|_{\infty} - \left\| \widehat{Q}^{\pi} - \widehat{Q}^{\star} \right\|_{\infty} - \left\| \widehat{Q}^{\pi^{\star}} - Q^{\star} \right\|_{\infty}.
\end{aligned}$$

$\rightsquigarrow$ The main technical problem: how to bound $Q^\pi - \widehat{Q}^\pi = \gamma(I - \gamma P^\pi)^{-1}(P - \widehat{P})\widehat{V}^\pi$?

$\rightsquigarrow$ Note that $\widehat{V}^\pi$ is a random variable that depends on $\widehat{P}$, therefore the standard concentration arguments cannot be applied.

$\rightsquigarrow$ This does not conflict with Simulation Lemma, where the evaluated policy is independent of the random process.

Crude Value Bounds

- Let's consider how to control the estimation error bounds in Lemma 2.

Lemma 3 (Crude Value Bouds[Azar et al., 2013]).

Let the failure probability $\delta \in (0,1)$. *With probability at least* $1-\delta$,

$$\left\| Q^\star - \widehat{Q}^{\pi^\star} \right\|_\infty \le \Delta_{\delta,N} \quad \text{and} \quad \left\| Q^\star - \widehat{Q}^\star \right\|_\infty \le \Delta_{\delta,N},$$

where

$$\Delta_{\delta,N} := \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}.$$

Proof of Lemma 3

For any policy π, we have

$$\begin{aligned}
Q^\pi - \widehat{Q}^\pi &= (I - \gamma P^\pi)^{-1} r - \left(I - \gamma \widehat{P}^\pi\right)^{-1} r \\
&= \left(I - \gamma \widehat{P}^\pi\right)^{-1} \left(\left(I - \gamma \widehat{P}^\pi\right) - (I - \gamma P^\pi)\right) Q^\pi \\
&= \gamma \left(I - \gamma \widehat{P}^\pi\right)^{-1} \left(P^\pi - \widehat{P}^\pi\right) Q^\pi \\
&= \gamma \left(I - \gamma \widehat{P}^\pi\right)^{-1} (P - \widehat{P}) V^\pi.
\end{aligned}$$

This bound is counterpart to the bound in Lemma 2 (c.f. Equation (2)).

Proof of Lemma 3

Let's consider $\pi^\star$, then we have

$$\left\|\gamma\left(I-\gamma\widehat{P}^{\pi}\right)^{-1}(\widehat{P}-P)V^\star\right\|_\infty \le \gamma\sum_{i=0}^{\infty}\left\|\gamma^i\left(\widehat{P}^{\pi}\right)^i(\widehat{P}-P)V^\star\right\|_\infty \le \gamma\sum_{i=0}^{\infty}\left\|\gamma^i(\widehat{P}-P)V^\star\right\|_\infty$$
$$\le \frac{\gamma}{(1-\gamma)}\cdot\sqrt{\frac{2\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N\cdot(1-\gamma)^2}}.$$

where we have used the Hoeffding's inequality that for a random variable X lies in $[a,b]$, consider N i.i.d. samples $X_i,\cdots,X_N$, then

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N}X_i-\mathbb{E}\left[X\right]\right|\le\sqrt{\frac{(b-a)^2}{2N}\log\left(\frac{2}{\delta}\right)}\right)\ge 1-\delta.$$

Proof of Lemma 3

For the second part, we have

$$\begin{aligned}
\left\|Q^\star - \widehat{Q}^\star\right\|_\infty &= \left\|\mathcal{T}Q^\star - \widehat{\mathcal{T}}\widehat{Q}^\star\right\|_\infty \\
&\le \left\|\mathcal{T}Q^\star - r - \widehat{P}^{\pi^\star}Q^\star\right\|_\infty + \left\|\widehat{P}^{\pi^\star}Q^\star + r - \widehat{\mathcal{T}}\widehat{Q}^\star\right\|_\infty \\
&= \gamma\left\|P^{\pi^\star}Q^\star - \widehat{P}^{\pi^\star}Q^\star\right\|_\infty + \gamma\left\|\widehat{P}^{\pi^\star}Q^\star - \widehat{P}^{\widehat{\pi}^\star}\widehat{Q}^\star\right\|_\infty \\
&= \gamma\left\|(P - \widehat{P})V^\star\right\|_\infty + \gamma\left\|\widehat{P}V^\star - \widehat{P}\widehat{V}^\star\right\|_\infty \\
&\le \gamma\left\|(P - \widehat{P})V^\star\right\|_\infty + \gamma\left\|V^\star - \widehat{V}^\star\right\|_\infty \\
&\le \gamma\left\|(P - \widehat{P})V^\star\right\|_\infty + \gamma\left\|Q^\star - \widehat{Q}^\star\right\|_\infty .
\end{aligned}$$

Therefore, we have $\left\|Q^\star - \widehat{Q}^\star\right\|_\infty \le \gamma/(1-\gamma)\left\|(P - \widehat{P})V^\star\right\|_\infty$.

Short Summary

- If we use the simple value bounds in Lemma 3 to Lemma 2, there exists an estimation error term with the order of $(1-\gamma)^{-4}$.
- To derive the tight bound, we cannot relax $(I-\gamma P^{\pi})$ as $(1-\gamma)^{-1}$ but use it to consider the variance of the estimator.

Important Lemma

Lemma 4.

For any policy and MDP M,

$$\left\| (I - \gamma P^{\pi})^{-1} \sqrt{\mathrm{Var}_P \left(V_M^{\pi} \right)} \right\|_{\infty} \leq \sqrt{\frac{2}{(1-\gamma)^3}}.$$

$\rightsquigarrow$ This bound is tight and is obtained by considering the Bellman style variance equation.

$\rightsquigarrow$ (In contrast, a naive derivation yields the order of $(1-\gamma)^{-2}$.)

Outline

Background and Notation

Main Result

Analysis and Proof

- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof

- Proof of Lemma 4
- Proof of Lemma 7

Motivation for Absorbing MDP

- To derive a tight bound, the authors claim that we need an understanding of quantities like $\left|(P - \widehat{P})\widehat{V}^\star\right|$ (and $\left|(P - \widehat{P})\widehat{V}^{\pi^\star}\right|$).
  - Originally, we need to consider $\left|(P - \widehat{P})\widehat{V}^{\pi}\right|$;
  - But it's (at least technically) connivent to consider $\left|(P - \widehat{P})\widehat{V}^\star\right|$ since we know there is only an optimization gap between $\widehat{V}^\star$ and $\widehat{V}^\pi$.
- However, we cannot directly apply a standard concentration argument because $\widehat{V}^\star$ (and $\widehat{V}^{\pi^\star}$) depends on $\widehat{P}$.
- To solve this issue, the authors introduce the absorbing MDPs where the dependence is decoupled by considering absorbing states.

Absorbing MDP

- Absorbing MDP $M_{s,u}$: $M_{s,u}$ is identical to M except that (only) the state s is absorbing in $M_{s,u}$, i.e., $P_{M_{s,u}}(s|s,a) = 1$ for all $a \in \mathcal{A}$; in addition, the reward at state s in $M_{s,u}$ is $(1-\gamma)u$, where u is a positive scalar.
- Notation: we use $V_{s,u}$ for the value function $V_{M_{s,u}}$ in $M_{s,u}$ and correspondingly for Q and reward and transition functions.
- By definition, we have that

$$V^{\pi}_{s,u}(s) = u. \tag{3}$$

- Similarly, we let $\widehat{M}_{s,u}$ denote the MDP uses the empirical model $\widehat{P}$ instead of P at all non-absorbing states.

Cover of Absorbing MDPs

- We can directly utilize the independence property to get the concentration bound on $\left|\left(P_{s,a} - \widehat{P}_{s,a}\right) \cdot \widehat{V}^{\star}_{s,u}\right|$.
- But let's think about the drawbacks of absorbing MDPs firstly.
  - The gap between $\widehat{V}^{\star}_{s,u}$ and $\widehat{V}^{\star}$? Since we care about $\left|\left(P_{s,a} - \widehat{P}_{s,a}\right) \cdot \widehat{V}^{\star}\right|$ ...
  - A worse bad news: $\widehat{V}^{\star}$ is a random variable, which implies that it's hard to exactly capture $\widehat{V}^{\star}$ with single absorbing MDP even though they could be close.
- Solution: let's consider the cover! (i.e., we use many absorbing MDPs)

Cover of Absorbing MDPs

- For some state s, we will consider $M_{s,u}$ for u in a finite set U_s, where

$$U_s \subset [V^\star(s) - \Delta_{\delta,N}, V^\star(s) + \Delta_{\delta,N}].$$

- In the following, we show that the concentration bound can be extended with a cover without additional high order terms.

Concentration at Absorbing State

Lemma 5.

Fix a state s, an action a, a finite set U_s and $\delta > 0$. With probability at least $1-\delta$, it holds for all $u \in U_s$:

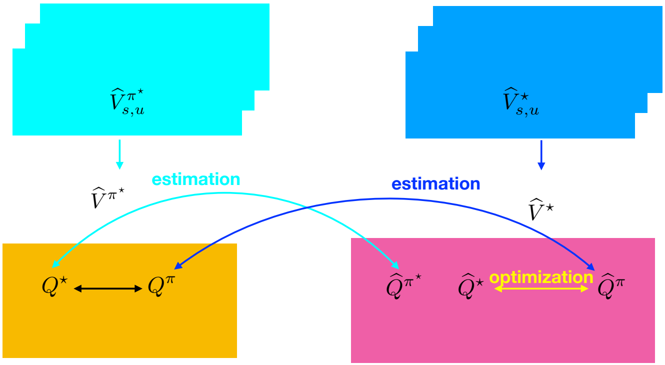
$$\left|\left(P_{s,a} - \widehat{P}_{s,a}\right) \cdot \widehat{V}^{\star}_{s,u}\right| \leq \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}\left(\widehat{V}^{\star}_{s,u}\right)} + \frac{2\log\left(4\left|U_s\right|/\delta\right)}{(1-\gamma)3N}$$

$$\left|\left(P_{s,a} - \widehat{P}_{s,a}\right) \cdot \widehat{V}^{\pi^\star}_{s,u}\right| \leq \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}\left(\widehat{V}^{\pi^\star}_{s,u}\right)} + \frac{2\log\left(4\left|U_s\right|/\delta\right)}{(1-\gamma)3N}$$

Proof.

Since $\widehat{P}_{s,a}$ and $\widehat{V}^{\star}_{s,u}$ are independent, the result directly comes from Bernstein's inequality along with a union bound over U_s. □

An Overview of The Proof



Remark on Lemma 5

- Note that the above argument heavily relies on the independence of $\widehat{P}_{s,a}$ and $\widehat{V}^\star_{s,u}$, which is not true for $\widehat{V}^\star$.
- Therefore, a natural question is: how to construct U_s such that for some $u \in U_s$, we have a good approximation of $\widehat{V}^\star$ based on $\widehat{V}^\star_{s,u}$.

Lemma 6.

Let $u^\star = V^\star_M(s)$ and $u^\pi = V^\pi_M(s)$. We have

$$V^\star_M(s) = V^\star_{s,u^\star}(s), \quad \forall s \in \mathcal{S}.$$

And for any deterministic policy π,

$$V^\pi_M(s) = V^\pi_{M_{s,u^\pi}}(s), \quad \forall s \in \mathcal{S}.$$

Proof of Lemma 6

- To prove the first claim, it suffices to show that $V_M^\star$ satisfies the Bellman optimality conditions on $M_{s,u^\star}$.
  - At state s, the Bellman optimality equations hold trivially.
  - For state $s' \neq s$, the outgoing transition model at s' in $M_{s,u^\star}$ is identical to that in M. Since $V_M^\star$ satisfies the Bellman optimality equations hold at every s' in M, it must hold for $M_{s,u^\star}$.
- The proof of the second claim is analogous.

Misspecification of u

- Lemma 6 provides a clue to select u but we also need robustness to misspecification of u.

Lemma 7.
For every state $s \in \mathcal{S}$, and $u, u' \in \mathbb{R}_+$ and any deterministic policy π,

$$\left\|Q^\star_{s,u} - Q^\star_{s,u'}\right\|_\infty \leq |u - u'| \quad \text{and} \quad \left\|Q^\pi_{s,u} - Q^\pi_{s,u'}\right\|_\infty \leq |u - u'|.$$

Note that $M_{s,u}$ and $M_{s,u'}$ are only different in the reward function at state s.

Outline

Background and Notation

Main Result

Analysis and Proof

- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof

- Proof of Lemma 4
- Proof of Lemma 7

Concentration on $\widehat{V}^\star$ and $\widehat{V}^{\pi^\star}$

Proposition 1.

Fix a state s, an action a, a finite set U_s, and $\delta > 0$. With probability at least $1 - 2\delta$, it holds that for all $u \in U_s$,

$$
\begin{aligned}
\left|\left(P_{s,a} - \widehat{P}_{s,a}\right) \cdot \widehat{V}^\star\right| \leq & \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}\left(\widehat{V}^\star\right)} \\
& + \min_{u \in U_s}\left|\widehat{V}^\star(s) - u\right| \left(1 + \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}}\right) + \frac{2\log\left(4\left|U_s\right|/\delta\right)}{(1-\gamma)3N} \\
\left|\left(P_{s,a} - \widehat{P}_{s,a}\right) \cdot \widehat{V}^{\pi^\star}\right| \leq & \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}} \sqrt{\mathrm{Var}_{P_{s,a}}\left(\widehat{V}^{\pi^\star}\right)} \\
& + \min_{u \in U_s}\left|\widehat{V}^{\pi^\star}(s) - u\right| \left(1 + \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}}\right) + \frac{2\log\left(4\left|U_s\right|/\delta\right)}{(1-\gamma)3N}
\end{aligned}
$$

Proof of Proposition 1

In Lemma 5, we have proved the error bounds of $\left|\left(P_{s,a}-\widehat{P}_{s,a}\right)\cdot\widehat{V}^{\star}_{s,u}\right|$. Based on this result, with probability at least $1-\delta$, for all $u\in U_s$, we have

$$
\begin{aligned}
&\left|\left(P_{s,a}-\widehat{P}_{s,a}\right)\cdot\widehat{V}^{\star}\right| \\
&=\left|\left(P_{s,a}-\widehat{P}_{s,a}\right)\cdot\left(\widehat{V}^{\star}-\widehat{V}^{\star}_{s,u}+\widehat{V}^{\star}_{s,u}\right)\right| \\
&\leq\left|\left(P_{s,a}-\widehat{P}_{s,a}\right)\cdot(\widehat{V}^{\star}-\widehat{V}^{\star}_{s,u})\right|+\left|\left(P_{s,a}-\widehat{P}_{s,a}\right)\cdot\widehat{V}^{\star}_{s,u}\right| \\
&\leq\left\|\widehat{V}^{\star}-\widehat{V}^{\star}_{s,u}\right\|_{\infty}+\sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}}\sqrt{\mathrm{Var}_{P_{s,a}}\left(\widehat{V}^{\star}_{s,u}\right)}+\frac{2\log\left(4\left|U_s\right|/\delta\right)}{(1-\gamma)3N} \\
&\leq\left\|\widehat{V}^{\star}-\widehat{V}^{\star}_{s,u}\right\|_{\infty}+\sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}}\sqrt{\mathrm{Var}_{P_{s,a}}\left(\widehat{V}^{\star}_{s,u}\right)}+\frac{2\log\left(4\left|U_s\right|/\delta\right)}{(1-\gamma)3N}.
\end{aligned}
$$

Proof of Proposition 1

Then, using the triangle inequality that $\sqrt{\mathrm{Var}_{P_{s,a}}(V_1 + V_2)} \le \sqrt{\mathrm{Var}_{P_{s,u}}(V_1)} + \sqrt{\mathrm{Var}_{P_{s,u}}(V_2)}$:

$$
\begin{aligned}
&\left|\left(P_{s,a} - \widehat{P}_{s,a}\right) \cdot \widehat{V}^\star\right| \\
\le &\left\|\widehat{V}^\star - \widehat{V}^\star_{s,u}\right\|_\infty + \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}}\sqrt{\mathrm{Var}_{P_{s,a}}\left(\widehat{V}^\star_{s,u} - \widehat{V}^\star + \widehat{V}^\star\right)} + \frac{2\log\left(4\left|U_s\right|/\delta\right)}{(1-\gamma)3N} \\
\le &\left\|\widehat{V}^\star - \widehat{V}^\star_{s,u}\right\|_\infty + \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}}\sqrt{\mathrm{Var}_{P_{s,a}}\left(\widehat{V}^\star_{s,u} - \widehat{V}^\star\right)} \\
&+ \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}}\sqrt{\mathrm{Var}_{P_{s,u}}(\widehat{V}^\star)} + \frac{2\log\left(4\left|U_s\right|/\delta\right)}{(1-\gamma)3N} \\
\le &\left\|\widehat{V}^\star - \widehat{V}^\star_{s,u}\right\|_\infty \left(1 + \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}}\right) + \sqrt{\frac{2\log\left(4\left|U_s\right|/\delta\right)}{N}}\sqrt{\mathrm{Var}_{P_{s,u}}(\widehat{V}^\star)} + \frac{2\log\left(4\left|U_s\right|/\delta\right)}{(1-\gamma)3N}
\end{aligned}
$$

Proof of Proposition 1

Finally, we note that the misspecification error between $\widehat{V}^\star$ and $\widehat{V}^\star_{s,u}$ is upper bounded by Lemma 6 and Lemma 7:

$$\left\|\widehat{V}^\star - \widehat{V}^\star_{s,u}\right\|_\infty = \left\|\widehat{V}^\star_{s,\widehat{V}^\star(s)} - \widehat{V}^\star_{s,u}\right\|_\infty \leq \left|\widehat{V}^\star(s) - u\right|.$$

Since the above bound holds for all $u \in U_s$, we may take the best possible choice, which completes the proof of the first claim. The proof of the second claim is analogous.

Bound With The Cover

Lemma 8.

With probability at least $1-\delta$,

$$\left|(P-\widehat{P})\widehat{V}^{\star}\right| \leq \sqrt{\frac{8L}{N}}\sqrt{\mathrm{Var}_P\left(\widehat{V}^{\star}\right)} + \Delta'_{\delta,N}\mathbf{1}$$

$$\left|(P-\widehat{P})\widehat{V}^{\pi^{\star}}\right| \leq \sqrt{\frac{8L}{N}}\sqrt{\mathrm{Var}_P\left(\widehat{V}^{\pi^{\star}}\right)} + \Delta'_{\delta,N}\mathbf{1},$$

where $L = \log\left(\frac{8|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\delta}\right)$, *and*

$$\Delta'_{\delta,N} = \sqrt{\frac{cL}{N}} + \frac{cL}{(1-\gamma)N}$$

with c being an absolute constant.

Proof of Lemma 8

$\rightsquigarrow$ We take U_s to be the evenly spaced elements in the interval $[V^\star - \Delta_{\delta/2,N}, V^\star + \Delta_{\delta/2,N}]$ and we take the size of U_s to be $|U_s| = \lceil (1-\gamma)^{-2} \rceil$.

$\rightsquigarrow$ By the crude value bounds in Lemma 3, with probability at least $1-\delta/2$, we have $\widehat{V}^\star \in [V^\star(s) - \Delta_{\delta/2,N}, V^\star(s) + \Delta_{\delta/2,N}]$ for all s.

$$
\begin{aligned}
\min_{u \in U_s} \left| \widehat{V}^\star(s) - u \right| &\leq \frac{2\Delta_{\delta/2,N}}{|U_s| - 1} \\
&= \frac{2}{|U_s| - 1} \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{4\log(4|\mathcal{S}||\mathcal{A}|/\delta)}{N}} \\
&\leq 4\gamma \sqrt{\frac{4\log(4|\mathcal{S}||\mathcal{A}|/\delta)}{N}}.
\end{aligned}
$$

Proof of Lemma 8

$\rightsquigarrow$ Now we use $\delta/(2|\mathcal{S}||\mathcal{A}|)$ so that the claims in Proposition 1 hold with probability greater than $1 - \delta/2$.

$\rightsquigarrow$ The first claim follows by substitution and noting that the probability of either event failing is less than $\delta/2$.

$\rightsquigarrow$ The proof of the second claim is analogous; note that Lemma 3 and Proposition 1 hold simultaneously so no further modification to the failure probability are required.

Outline

Background and Notation

Main Result

Analysis and Proof

- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof

- Proof of Lemma 4
- Proof of Lemma 7

The Estimation Error Bound

$\rightsquigarrow$ We present the last lemma and show that Theorem 1 follows from this lemma.

Lemma 9.

Let $\widehat{\pi}$ be the output of MBRL algorithm. Then with probability at least $1-\delta$, we have

$$\left\|Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}}\right\|_\infty \le \frac{\gamma}{1-\alpha_{\delta,N}}\left(\sqrt{\frac{c}{(1-\gamma)^3}LN} + \frac{cL}{(1-\gamma)^2 N}\right) + \frac{1}{1-\alpha_{\delta,N}} \cdot \frac{\gamma \epsilon_{\mathrm{opt}}}{1-\gamma}\left(1+\sqrt{\frac{L}{N}}\right)$$

$$\left\|Q^{\star} - \widehat{Q}^{\pi^\star}\right\|_\infty \le \frac{\gamma}{1-\alpha_{\delta,N}}\left(\sqrt{\frac{c}{(1-\gamma)^3}LN} + \frac{cL}{(1-\gamma)^2 N}\right).$$

where c is an absolute constant and where $\alpha_{\delta,N} = \gamma/(1-\gamma)\sqrt{8L/N}$.

Proof of Theorem 1

By Lemma 2, we have

$$Q^{\widehat{\pi}} \geq Q^{\star} - \left\| Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}} \right\|_{\infty} - \epsilon_{\mathrm{opt}} - \left\| \widehat{Q}^{\pi^{\star}} - Q^{\star} \right\|_{\infty}.$$

By the choice of $N \gtrsim (1-\gamma)^{-3}\epsilon^{-2}L$, we have $\alpha_{\delta,N} = \frac{\gamma}{1-\gamma}\sqrt{\frac{8L}{N}} \leq \frac{1}{2}$. This and Lemma 9 implies:

$$Q^{\widehat{\pi}} \geq Q^{\star} - 4\gamma\left(\sqrt{\frac{c}{(1-\gamma)^3} \cdot \frac{L}{N}} + \frac{c \cdot L}{(1-\gamma)^2 N}\right) - \frac{4\gamma\epsilon_{\mathrm{opt}}}{1-\gamma} - \epsilon_{\mathrm{opt}}.$$

Plugging in the choice of N yields an ϵ-optimal policy as desired.

Proof of Lemma 9

We have that

$$
\begin{aligned}
\left\| Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}} \right\|_\infty &\overset{(a)}{=} \gamma \left\| (I - \gamma P^{\widehat{\pi}})^{-1} (P - \widehat{P}) \widehat{V}^{\widehat{\pi}} \right\|_\infty \\
&\overset{(b)}{\leq} \gamma \left\| (I - \gamma P^{\widehat{\pi}})^{-1} (P - \widehat{P}) \widehat{V}^{\star} \right\|_\infty + \gamma \left\| (I - \gamma P^{\pi})^{-1} (P - \widehat{P}) (\widehat{V}^{\widehat{\pi}} - \widehat{V}^{\star}) \right\|_\infty \\
&\overset{(c)}{\leq} \gamma \left\| (I - \gamma P^{\widehat{\pi}})^{-1} (P - \widehat{P}) \widehat{V}^{\star} \right\|_\infty + \frac{\gamma \epsilon_{\mathrm{opt}}}{1 - \gamma} \\
&\overset{(d)}{\leq} \gamma \left\| (I - \gamma P^{\widehat{\pi}})^{-1} \left| (P - \widehat{P}) \widehat{V}^{\star} \right| \right\|_\infty + \frac{\gamma \epsilon_{\mathrm{opt}}}{1 - \gamma}
\end{aligned}
$$

where (a) uses Lemma 2; (b) is based on triangle inequality; (c) is based on the fact that for a vector $v \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, $\left\| (I - \gamma P^{\pi})^{-1} v \right\|_\infty \leq (1 - \gamma)^{-1} \left\| v \right\|_\infty$; (d) uses that $(I - \gamma P^{\widehat{\pi}})$ has all positive entries.

Proof of Lemma 9

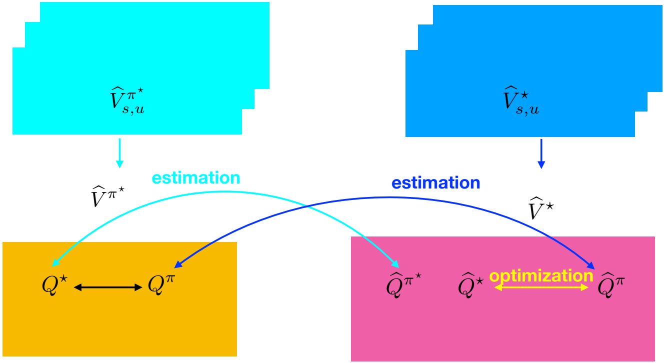
$$
\begin{aligned}
&\left\| Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}} \right\|_\infty \\
&\overset{(e)}{\leq} \gamma \sqrt{\frac{8L}{N}} \left\| (I - \gamma P^{\widehat{\pi}})^{-1} \sqrt{\mathrm{Var}_P(\widehat{V}^\star)} \right\|_\infty + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\mathrm{opt}}}{1-\gamma} \\
&\leq \gamma \sqrt{\frac{8L}{N}} \left\| (I - \gamma P^{\widehat{\pi}})^{-1} \left(\sqrt{\mathrm{Var}_P(V^{\widehat{\pi}})} + \sqrt{\mathrm{Var}_P(V^{\widehat{\pi}} - \widehat{V}^{\widehat{\pi}})} + \sqrt{\mathrm{Var}_P(\widehat{V}^{\widehat{\pi}} - \widehat{V}^\star))} \right) \right\|_\infty \\
&\quad + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\mathrm{opt}}}{1-\gamma} \\
&\overset{(g)}{\leq} \gamma \sqrt{\frac{8L}{N}} \left(\sqrt{\frac{2}{(1-\gamma)^3}} + \frac{\sqrt{\left\| V^{\widehat{\pi}} - \widehat{V}^{\widehat{\pi}} \right\|_\infty^2}}{1-\gamma} + \frac{\epsilon_{\mathrm{opt}}}{1-\gamma} \right) + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\mathrm{opt}}}{1-\gamma}
\end{aligned}
$$

Proof of Lemma 9

$$
\begin{aligned}
&\left\| Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}} \right\|_\infty \\
&\leq \gamma \sqrt{\frac{8L}{N}} \left(\sqrt{\frac{2}{(1-\gamma)^3}} + \frac{\left\| Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}} \right\|_\infty}{1-\gamma} + \frac{\epsilon_{\text{opt}}}{1-\gamma} \right) + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\text{opt}}}{1-\gamma} \\
&= \gamma \sqrt{\frac{8L}{N}} \left(\sqrt{\frac{2}{(1-\gamma)^3}} + \frac{\left\| Q^{\widehat{\pi}} - \widehat{Q}^{\widehat{\pi}} \right\|_\infty}{1-\gamma} \right) + \frac{\gamma \Delta'_{\delta,N}}{1-\gamma} + \frac{\gamma \epsilon_{\text{opt}}}{1-\gamma} \left(1 + \sqrt{\frac{8L}{N}} \right).
\end{aligned}
$$

where (e) uses Lemma 8 and (f) applies Lemma 4.

Summary



$\widehat{V}^{\pi^\star}_{s,u}$
$\widehat{V}^\star_{s,u}$
$\widehat{V}^{\pi^\star}$
estimation
estimation
$\widehat{V}^\star$
$Q^\star$
Q^π
$\widehat{Q}^{\pi^\star}$
$\widehat{Q}^\star$
optimization
$\widehat{Q}^\pi$

Outline

Background and Notation

Main Result

Analysis and Proof

- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof

- Proof of Lemma 4
- Proof of Lemma 7

Outline

Background and Notation

Main Result

Analysis and Proof
- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof
- Proof of Lemma 4
- Proof of Lemma 7

Proof of Lemma 4

$\rightsquigarrow$ Note that $(1-\gamma)(I-\gamma P^\pi)^{-1}$ is matrix that each row is a probability distribution.

$\rightsquigarrow$ For a positive vector v, Jensen's inequality suggests that $\mathbb{E}\left[v\right] \leq \sqrt{\mathbb{E}\left[v\right]}$.

$$
\begin{aligned}
\left\| \left(I-\gamma P^\pi\right)^{-1}\sqrt{v} \right\|_\infty &= \frac{1}{1-\gamma}\left\| (1-\gamma)\left(I-\gamma P^\pi\right)^{-1}\sqrt{v} \right\|_\infty \\
&\leq \sqrt{\left\| \frac{1}{1-\gamma}\left(I-\gamma P^\pi\right)^{-1} v \right\|_\infty} \\
&\leq \sqrt{\left\| \frac{2}{1-\gamma}\left(I-\gamma^2 P^\pi\right)^{-1} v \right\|_\infty},
\end{aligned}
$$

where the last line is based on $\left\| I-\gamma P^\pi v \right\|_\infty \leq 2\left\| \left(I-\gamma^2 P^\pi\right)^{-1} v \right\|_\infty$ (which we will prove later).

Proof of Lemma 4

Our main proof is completed as follows: by Equation (1),

$$\Sigma_M^\pi = \gamma^2 (I - \gamma P^\pi)^{-1} \mathrm{Var}_P(V_M^\pi).$$

Then, take $v = \mathrm{Var}_P(V_M^\pi)$, we have that

$$\begin{aligned}
\left\| (I - \gamma P^\pi)^{-1} \sqrt{\mathrm{Var}_P(V_M^\pi)} \right\|_\infty &\leq \sqrt{\left\| \frac{2}{1-\gamma} \left(I - \gamma^2 P^\pi\right)^{-1} \mathrm{Var}_P(V_M^\pi) \right\|_\infty} \\
&\leq \sqrt{\frac{2}{1-\gamma} \frac{\gamma^2}{(1-\gamma)^2}},
\end{aligned}$$

where we note that $\Sigma_M^\pi \leq \gamma^2/(1-\gamma)^2$ by definition.

Proof of Lemma 4

Let's prove $\|I-\gamma P^{\pi} v\|_{\infty} \leq 2\left\|\left(I-\gamma^2 P^{\pi}\right)^{-1} v\right\|_{\infty}$ now.

$$
\begin{aligned}
\left\|\left(I-\gamma P^{\pi}\right)^{-1} v\right\|_{\infty} & =\left\|\left(I-\gamma P^{\pi}\right)^{-1}\left(I-\gamma^2 P^{\pi}\right)\left(I-\gamma^2 P^{\pi}\right)^{-1} v\right\|_{\infty} \\
& =\left\|\left(I-\gamma P^{\pi}\right)^{-1}\left((1-\gamma) I+\gamma\left(I-\gamma P^{\pi}\right)\right)\left(I-\gamma^2 P^{\pi}\right)^{-1} v\right\|_{\infty} \\
& =\left\|\left((1-\gamma)\left(I-\gamma P^{\pi}\right)^{-1}+\gamma I\right)\left(I-\gamma^2 P^{\pi}\right)^{-1} v\right\|_{\infty} \\
& \leq(1-\gamma)\left\|\left(I-\gamma P^{\pi}\right)^{-1}\left(I-\gamma^2 P^{\pi}\right)^{-1} v\right\|_{\infty}+\gamma\left\|\left(I-\gamma^2 P^{\pi}\right)^{-1} v\right\|_{\infty} \\
& \leq \frac{1-\gamma}{1-\gamma}\left\|\left(I-\gamma^2 P^{\pi}\right)^{-1} v\right\|_{\infty}+\gamma\left\|\left(I-\gamma^2 P^{\pi}\right)^{-1} v\right\|_{\infty} \\
& \leq 2\left\|\left(I-\gamma^2 P^{\pi}\right)^{-1} v\right\|_{\infty} .
\end{aligned}
$$

Outline

Background and Notation

Main Result

Analysis and Proof
- Errors in Empirical Estimates
- An s-absorbing MDP
- Key Analysis
- Proof of Main Theorem

Additional Proof
- Proof of Lemma 4
- Proof of Lemma 7

Proof of Lemma 7

We observe that the reward functions are different only at state s, thus

$$\left\|r_{s,u} - r_{s,u'}\right\|_\infty = (1-\gamma)\left|u - u'\right|.$$

Let $\pi_{s,u}$ be the optimal policy in $M_{s,u}$:

$$\begin{aligned} Q^\star_{s,u} - Q^\star_{s,u'} &= Q^\star_{s,u} - \max_\pi (I - \gamma P^\pi_{s,u'})^{-1} r_{s,u'} \leq Q^\star_{s,u} - (I - \gamma P^{\pi_{s,u}}_{s,u'}) r_{s,u'} \\ &= (I - \gamma P^{\pi_{s,u}}_{s,u'})^{-1}(r_{s,u} - r_{s,u'}) \leq \frac{1}{1-\gamma}\left\|r_{s,u'} - r_{s,u}\right\|_\infty \\ &= \left|u - u'\right|. \end{aligned}$$

The proof of the lower bound is analogous; the proof of the second argument can be obtained in a similar way.

References I

A. Agarwal, S. M. Kakade, and L. F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In Annual Conference on Learning Theory, pages 67–83, 2020.

M. G. Azar, R. Munos, and H. J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. Machine Learning, 91(3):325–349, 2013.